Enhanced Video Surveillance Systems for “Signal For Help” Detection on Edge Devices

Enhanced Video Surveillance Systems for "Signal For Help" Detection on Edge Devices

Sarah Azimi
Department of Computer and Control Engineering
Politecnico di Torino
Turin, Italy
sarah.azimi@polito.it

Corrado De Sio
Department of Computer and Control Engineering
Politecnico di Torino
Turin, Italy
corrado.desio@polito.it

Luca Sterpone
Department of Computer and Control Engineering
Politecnico di Torino
Turin, Italy
luca.sterpone@polito.it

***Abstract*— The COVID-19 pandemic triggered a concerning rise in violence against women and children, known as The Shadow Pandemic. To address this, a Canadian foundation introduced the "Signal for Help" gesture to discreetly alert others in danger. However, the effectiveness of this approach depends on individuals recognizing and responding to the signal. In this paper, we propose an innovative solution that adopts the technology available in smart cities to detect the "Signal for Help" in real-time through surveillance footage. We developed and implemented a recognition algorithm on an affordable device that achieves accurate detection of the signal in 94% of cases. This approach has the potential to improve the response to instances of violence, providing a reliable means of alerting authorities and support networks.**

Keywords—CNN, Violence Recognition, Edge Devices

I. INTRODUCTION

The intersection of technology and society has given rise to innovative solutions to societal issues. The tragic story of a 35-year-old woman in Rome serves as a stark reminder of the importance of using technology to help individuals in need [1] The woman was at dinner with her ex-partner when she felt unsafe and attempted to use a "signal for help". Unfortunately, nobody recognized the signal, and she was forced to leave the restaurant with her ex-partner, who later shot and killed her.

But what was the signal that the women tried to use? The signal or the hand gestures is known as "Signal for Help", represented in Figure 1. It is a hand signal that can be used to silently communicate to others that you need help and are in danger. The signal was first created by the Canadian Women's Foundation during the COVID-19 pandemic in response to increasing rates of domestic violence. Since then, it gained global recognition and has been used to provide a discreet way for victims of abuse to communicate their need for help, as the young women in a restaurant in Rome tried to do.

This incident raises important questions about how technology can be used to prevent such tragedies from happening in the future. Smart cities, with their advanced infrastructure and connectivity, present a unique opportunity to leverage technology to create safer communities. By adopting technologies such as IoT devices and AI-powered surveillance systems, we can create a system that can automatically recognize and respond to emergency signals, including the "Signal for Help."

In this paper, we propose a system that utilizes smart city technology to recognize emergency signals and alert the appropriate authorities. By integrating surveillance cameras, and AI algorithms, the system can detect the "Signal for Help", and alert authorities in real-time. This system could be implemented in public spaces such as restaurants and malls, providing individuals with a discreet way to signal for help and increasing the chances of receiving assistance.

Fig. 1. The "Signal for Help" hand gesture.

II. RELATED WORKS

The impact of emotions on human life is undeniable, as they are considered one of the most fundamental means of communication. As a result, extensive research efforts are dedicated to developing techniques for automatic emotion recognition [2]. One specific area of interest is the detection of violence in videos, where computer vision methods, including object and motion detection and classification, are employed [3].

Due to the complexity and diversity of violence patterns, many approaches have been developed. The work in [4] proposes a model to detect fights based on extracting motion acceleration while achieving 78% accuracy in identifying fights. In [5], a model is proposed consisting of a series of convolutional layers for extracting discriminable features while attaining an accuracy of around 97% in classifying violent and non-violent videos. Additionally, some studies focus on violence detection by analyzing skin and blood regions for rapid motions [6]. However, defining effective features for violence detection remains challenging due to variations in the human body and the wide range of violence types, leading to a high rate of false detection.

In our specific case, discreetly recognizing the "Signal for Help" hand gesture, conventional methods for detecting violence-related characteristics are not applicable. Thus, it becomes necessary to explore specialized platforms that excel in hand gesture detection and recognition. Hand gestures serve as crucial nonverbal communication in Human-Machine Interaction (HMI). Recent research endeavors have been dedicated to hand gesture recognition, addressing tasks such as gesture control in vehicles, sports referee signals, and sign languages [7][8]. Convolutional Neural Networks (CNNs) have demonstrated remarkable success in these domains [9].

979-8-3503-2486-0/23/$31.00 ©2023 IEEE

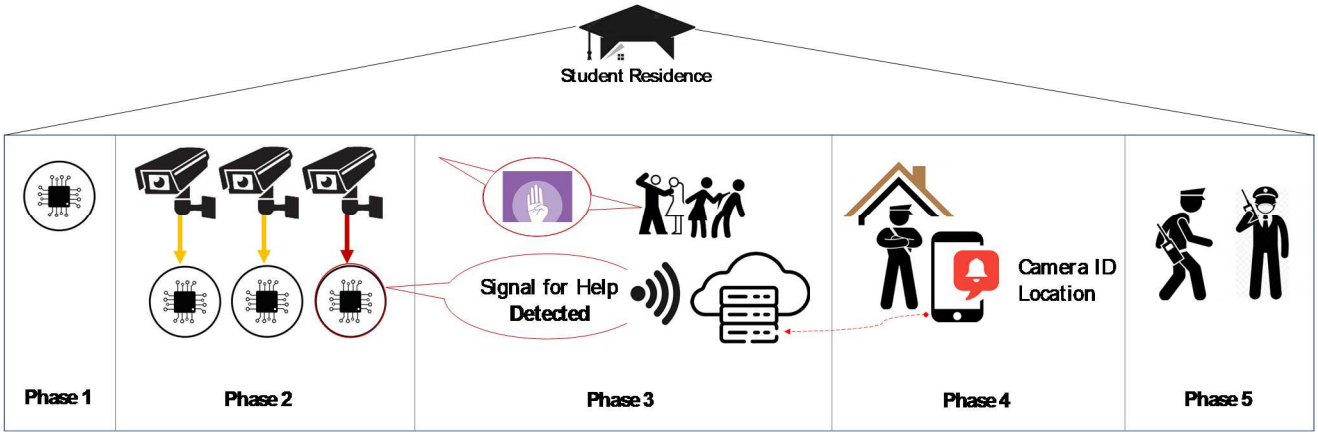

Fig. 2. Phases of the "Signal For Help" Recognition Platform: Phase 1 – Development of the recognition algorithm and implementation on edge device, Phase 2- Integration of the developed platform with surveillance cameras, Phase 3- Communication of Recognition platform with Cloud services upon detection of "Signal for Help", Phase 4- Sending a notification to the cellphone of the guard, Phase 5- Informing right authorities for further help.

While significant attention has been given to studying static hand gestures, which involve fixed hand poses, dynamic hand gesture recognition, where finger shapes have received relatively less focus. While many research works address detection and classification as separate tasks, real-time hand gesture recognition requires the simultaneous application of both to continuous video streams. A proposed architecture employs two CNN models: ResNet for hand detection and ResNext as an offline working classifier, achieving a classification accuracy of 94.04% on publicly available datasets [10]. However, it is not suitable for implementation on resource-limited embedded systems. For the platform to be widely used in public places such as restaurants, supermarkets, and building entrances, it must not only be highly accurate but also cost-effective.

A. The Main Contribution

This paper aims to develop a low-cost platform that can perform real-time "Signal for Help" hand gesture recognition through surveillance videos adopting commercial-of-the-shelf edge devices. The main contribution of this paper is the development of a platform that can perform real-time hand gesture recognition for "Signal for Help" using surveillance videos. The platform is designed to require low computational resources, enabling it to be implemented on edge devices and embedded systems. This is a significant advantage over expensive GPU-based systems that provide higher performance [14] but come at a much higher cost. The recognition algorithms are based on two well-known CNN architectures, MediaPipe and MobileNet, for performing recognition tasks. The developed platform is implemented on the Xilinx Ultra96-V2 edge board which hosts an embedded Arm processor. Exploiting the open-source "Signal for Help" test dataset shows 94% of true detection which confirms the efficiency of the proposed method.

III. OBJECTIVE

While the "Signal for Help" has proven to be an effective way for individuals to silently communicate the occurrence of violence, its success ultimately depends on others recognizing the signal or knowing its meaning.

To overcome this challenge, we proposed a solution that is not dependent on the human being for recognition of "Signal for Help", but dependent on the technology in smart cities that are becoming more and more advanced every day. To elaborate more, we developed a comprehensive workflow that utilizes real-time detection of the signal through Video Camera Recording (VCR) surveillance videos. By alerting the appropriate authorities, we can ensure prompt assistance and follow-up in the event of an emergency. This solution aims to enhance the effectiveness of the "Signal for Help" and increase the likelihood of individuals receiving the help they need.

The workflow comprises five phases:

1. Development of a real-time gesture recognition platform and implementation it on a low-cost edge device yet, high performance.
2. Integrating the developed platform with the already existing surveillance videos across smart cities.
3. Establishing a communication mechanism between the developed platform and the cloud server to store the occurrence of the event.
4. Sending a push notification to the cellphone of the nearest authority such as a building guard or supermarket staff.
5. Forwarding the notification to the appropriate authorities to request further actions.

Figure 2 represents the mapping of different phases of the proposed workflow considering a student residence. While we are working simultaneously on different phases of this project, this paper is dedicated to elaborating on the preliminary methodology and results obtained on the development and implementation of the gesture recognition algorithm.

IV. METHODOLOGY

We have developed a platform that leverages smart city technologies to enhance social safety. The platform can detect and recognize the real-time "Signal for Help" hand gesture through surveillance videos, and it is optimized for low computational resources. To achieve this, we utilized two CNN architectures, MediaPipe and MobileNet, to create a single, high-performance, and lightweight hierarchical recognition architecture. The platform operates in two steps: hand detection and hand gesture classification.

The MediaPipe framework is used for the hand detection step, checking for the existence of hands in video frames [12]. If one or more hands are detected, they are passed on to the next step. The MediaPipe output provides a binary signal to notify the classifier that hand detection is active. The classifier

is based on the MobileNet model [11], which checks for the presence of hand gestures. If detection is triggered, the platform sends and stores the notification on the server through the hardware platform's network connection.

A. The Hand Detection

The first step of our framework is dedicated to detecting hands in videos using the MediaPipe platform. Developed by Google, MediaPipe Hand is a pre-trained model included with the framework that tracks hand palms and fingers.

MediaPipe Hand consists of two CNN models, one for palm detection and the other for hand landmark recognition. The palm detector identifies the presence of hands in an image and outputs their bounding box. The hand landmark model then provides the key points of the hand.

Our approach uses MediaPipe Hand to detect when one or more hands are present in the video, enabling the next step of the platform. Each frame of the video is evaluated individually and forwarded to the next step. Since MediaPipe is small, lightweight, and efficient, it can be implemented on embedded IoT devices like mobile phones or smart cameras without specific resource tailoring.

B. The Hand Gesture Recognition

The second stage of our platform builds upon the hand detection results from the first stage. To ensure compatibility between the two stages, each frame is scaled and the videos are down-sampled. Time-consecutive frames are then accumulated and provided to the second phase, which utilizes the MobileNet Model developed by Google.

MobileNet is specifically designed for mobile vision applications, including classification tasks, and is optimized for lightweight systems. The second stage evaluates whether the hand gesture appears in the video. The input to this stage is a collection of time-consecutive frames, and the output is a binary signal indicating the presence or absence of the "Signal for Help" gesture.

C. Training the Platform

Please notice that MediaPipe is already trained and provides an accuracy of more than 95%, therefore, no additional training has been needed. However, MobileNet required to be trained with a dataset to classify the “Signal For Help”. To do so, we used the open-source dataset, available online [14]. However, due to the small size of the available dataset, we first, exploit the open-source hand gesture dataset, the Jester dataset which includes 27 hand gestures [13], and then, we performed a fine-tunning phase in which the last year of the network is trained and tuned with the small “Signal for Help” dataset [14]. Please note that the videos available in the “Signal for Help” dataset are featured in a way to provide realistic scenarios. To elaborate more, the videos are collected considering 0 to 4/5 m distance from the camera, with indoor and outdoor lights, in the presence of single/multiple people/hands in the videos.

We trained the MobileNet model with the Jester dataset for 40 epochs, on a machine with an NVIDIA GeForce RTX 2080 Ti GPU, which required 26 hours. We have used an SGD optimizer with a learning rate that started at 0.1 and has been divided by 10 at the 15th, 25th, and 35th epochs, cross-entropy loss function, a batch size of 64, a dampening of 0.9, and weight decay at 0.001.

TABLE 1. TABLE TYPE STYLES

Parameters	***Jesture (Training)***	***Signal for Help (Fine Tunning)***
Dataset Size	148,092	200
Classes of Dataset [#]	27	2
Epochs [#]	39	39
Training Time [h]	26	2
Accuracy [%]	91.71	91.25

V. RESULTS AND DISCUSSION

Our platform has been successfully implemented on a low-cost System-on-Chip development board, specifically the Xilinx Ultra96-V2. This section presents the results of our implementation, highlighting the accuracy and high performance of the developed platform.

Despite the limited resources of the low-cost hardware, our platform is designed to be lightweight and optimized for low computational resources. As a result, it is suitable for implementation on devices with limited resources, such as the Ultra96-V2.

A. Hardware Implementation Characteristics

The board chosen for the implementation of the platform is Xilinx Ultra96-V2. It is a low-cost, versatile System-on-Chip (SoC) development board designed for a wide range of applications, including artificial intelligence (AI), Internet-of-Things (IoT), and embedded vision applications.

The board features a Xilinx Zynq UltraScale+ MPSoC device, which integrates an Arm processor, and programmable logic. The operating system running on the Arm processor in the Zynq SoC is typically Linux, which provides a robust and mature platform for development. Moreover, a Python overlay for the Linux operating system running on the Zynq SoC is provided which has been exploited to implement and execute the developed platforms.

B. Verification of “Signal for Help” Recognition Platform

After completing the training phase, we proceeded to verify the effectiveness of our developed workflow for real-time detection of the "Signal for Help" gesture. To accomplish this, we divided the open-source "Signal for Help" dataset into two sections, validation, and test, and used the available test set to execute and evaluate our proposed framework.

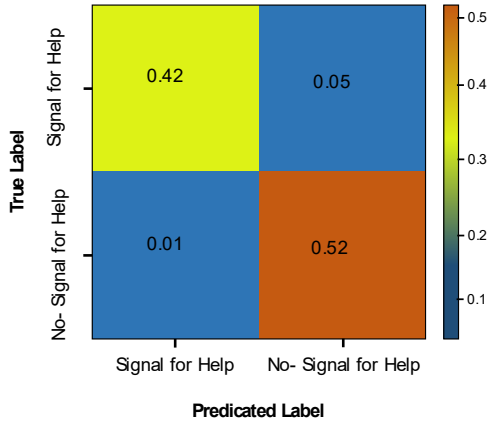

Fig. 3. Confusion Matrix for the developed Platform with the “Signal for Help” dataset.

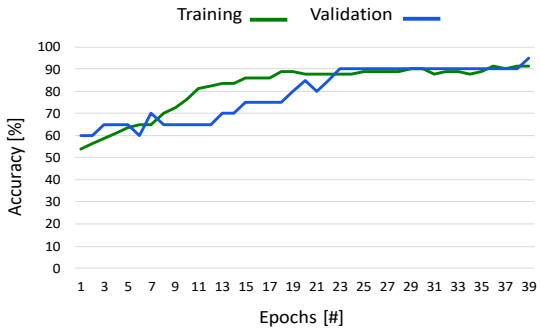

Fig. 4. Accuracy obtained after different epochs for MobileNet file-tuned with the "Signal for Help" Dataset.

Our recognition platform operates at an average rate of 24 fps when no gesture is detected, with only the MediaPipe model active for detecting the presence of the hand. However, when both the MediaPipe and MobileNet are active, performing detection of the hand's presence and recognition of the "Signal for Help" gesture, it runs at a rate of 5 fps.

We tested the videos in the test set of the dataset and found that our platform successfully detected a true alarm in 94% of the cases. We also observed a 5% false alarm rate and a 1% missing alarm rate, as shown in Figure 3, which displays the confusion matrix of our developed platform using the "Signal for Help" test set and Figure 4 shows the accuracy obtained after different epochs of MobileNet fine-tuned with the "Signal for Help" Dataset. Additionally, Figure 5 showcases some examples of real-time "Signal for Help" gesture detection through video.

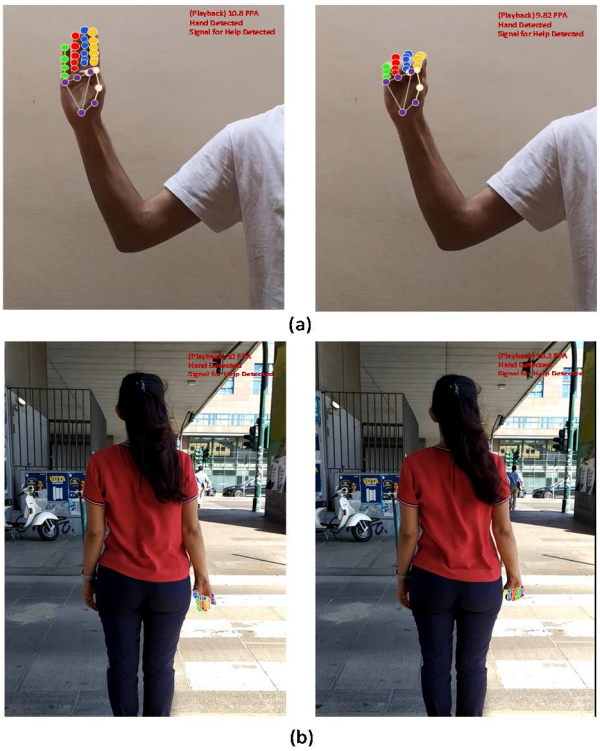

Fig. 5. Executing the Developed Platform on the test set of the "Signal for Help" Dataset: (a) close distance from the camera with the plane background (b) far distance from the camera with the urban background.

VI. Conclusions

This paper is dedicated to the introduction of a comprehensive workflow, for real-time detection of "Signal for Help" through surveillance videos, starting from the development and implementation of the "Signal for Help" recognition algorithm to informing the right authorities. While different phases of the workflow are being developed, this paper is elaborating on one of the most important phases of the workflow, the development of a real-time recognition algorithm and implementing it on an economic edge device. The developed framework is executed on an open-source "Signal for Help" dataset which confirms its efficiency. Moreover, currently, we are working on the development of a mobile application that is capable to receive a push notification as soon as one "Signal for Help" event occurred together with information such as the location, date/time of occurrence, and other information. The mentioned platform can be installed by the right authorities such as guards of the banks or supermarkets to help as soon as possible.

References

[1] https://roma.repubblica.it/cronaca/2023/01/14/news/donna_uccisa_da_un_uomo_tuscolano-383468200/.

[2] Kiruthiga P and R. Rajavel, "Audio Visual Emotion Recognition in Children", International Conference on Power of Digital Technologies in Societal Empowerment, CHENCON2021, 2021.

[3] M. Ramzen et al., "A Review on State-of-the-art Violence Detection Techniques", in EEE Access, vol. 7, pp. 107560-107575, 2019, DOI: 10.1109/ACCESS.2019.2932114.

[4] E. Y. Fu, H. Va Leong, G. Ngai, and S. Chan, "Automatic Fight Detection in Surveillance Videos," in Proceedings of the 14th International Conference on Advances in Mobile Computing and Multi-Media, ser. MoMM '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 225–234, doi: 10.1145/3007120.3007129.

[5] S. Sudhakaran and O. Lanz, "Learning to detect violent videos using convolutional long short-term memory," 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1-6, doi: 10.1109/AVSS.2017.8078468.

[6] L.-H. Chen, H.-W. Hsu, L.-Y. Wang, and C.-W. Su. "Violence detection in movies" In International conference on Computer Graphics, Imaging and Visualization (CGIV), 2011.

[7] K. A. Smith, C. Csech, D. Murdoch and G. Shaker, "Gesture Recognition Using mm-Wave Sensor for Human-Car Interface," in IEEE Sensors Letters, vol. 2, no. 2, pp. 1-4, June 2018, Art no. 3500904, doi: 10.1109/LSENS.2018.2810093.

[8] Ashwini, R Amutha, R Rajavel, D Anusha, "Classification of Daily Human Activities Using Wearable Inertial Sensor", 2020 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), pp 1-6, 2020.

[9] Rajangam Athilakshmi, Ramadoss Rajavel, Shomona Gracia Jacob, "A survey on deep-learning architectures", Journal of Computational and Theoretical Nanoscience, Volume 15, Issue 8, Pages 2577-2579, 2018.

[10] O. Köpüklü, A. Gunduz, N. Kose and G. Rigoll, "Real-time Hand Gesture Detection and Classification Using Convolutional Neural Networks," 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 2019, pp. 1-8, doi: 10.1109/FG.2019.8756576.

[11] A. G. Howard, et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications", in Computer Vision and Pattern Recognition, 2017, arXiv:1704.04861.

[12] Okan Kopuklu, Neslihan Kose, Ahmet Cunduz, Gerhard Rigoll, "Resource Efficient 3D Convolutional Neural Networks", in International Conference on Computer Vision, 2019.

[13] J. Materzynska, G. Berger, I. Bax and R. Memisevic, "The Jester Dataset: A Large-Scale Video Dataset of Human Gestures," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 2874-2882, doi: 10.1109/ICCVW.2019.00349.

[14] S. Azimi, et al., "Fighting for a future free from violence: A framework for real-time detection of "Signal for Help"", in Intelligent Systems with Applications, Vol. 17, 2023, DOI: 10.1016/j.iswa.2022.200174.